一种改进 K-means 聚类的 FCMM 算法 *

杨明极，马　池，王　娅，张　竹

(哈尔滨理工大学 测控技术与通信工程学院, 哈尔滨 150080)

摘　要：针对 K-means 算法易受初始聚类中心影响而陷入局部最优的问题，提出一种基于萤火虫智能优化和混沌理论的 FCMM 算法。首先利用最大最小距离算法确定聚类类别值 K 和初始聚类中心位置；然后以各聚类中心为基准点，利用 Tent 映射构建混沌空间，通过混沌搜索更新聚类中心，以降低初始聚类中心过于临近的影响，并改善算法易陷入局部最优的问题。仿真结果表明，FCMM 算法的平均聚类精度相较于经典 K-means 算法和 FA 算法分别提高了 7.51%和 2.2%，成功避免算法陷入局部最优解，提高了划分初始数据集的效率和寻优精度。

关键词：K-means 聚类；萤火虫；最大最小距离；Tent 映射；混沌搜索

中图分类号：TP301.6　　doi: 10.3969/j.issn.1001-3695.2017.12.0851

Algorithm named FCMM to improve K-means clustering algorithm

Yang Mingji, Ma Chi, Wang Ya, Zhang Zhu

(*School of Measure-control Technology & Communication Engineering*, *Harbin University of Science & Technology*, *Harbin* 150080, *China*)

Abstract: In order to solve the problem that the K-means algorithm gets affected by the initial cluster centers easily, this paper proposes FCMM algorithm based on firefly intelligence optimization and chaos theory. It uses the max-min distance clustering algorithm to calculate the number K of cluster center and determine the location of initial cluster centers. To overcome the problem that initial clustering centers are too close to each other and traditional algorithm falls into local optima easily, it uses Tent mapping to construct a chaotic space with each cluster center as the datum point, and then updates cluster centers through chaotic search. The experimental results show that the average clustering accuracy of the FCMM algorithm than that of the classical K-means algorithm and the FA algorithm is respectively 7.51% and 2.2% higher, the FCMM algorithm avoids falling into the local optimal solution successfully, and improves the efficiency and precision of the initial data set.

Key words: K-means clustering; firefly; maximum and minimum distance; Tent mapping; chaotic search

0　引言

伴随大数据时代的到来，数据挖掘领域得到日新月异的发展。聚类分析作为数据挖掘和数据分析的经典方法[1]，在模式识别，室内定位，统计学等领域拥有良好的发展前景。随着社会发展对数据精度要求的不断提高，如何提高算法的聚类精度是广大学者一直以来的研究热点[2]。

鉴于传统 K-means 聚类算法存在易受初始聚类中心和异常数据影响的缺陷[3]，算法开始研究集中于利用特征关联度对传统 K-means 算法的初始聚类中心进行优化[4]和基于自适应权重的聚类算法[5]。随着智能优化算法的提出和发展，萤火虫算法 FA(Firefly Algorithm)作为一种源于生物界的随机优化算法，因其具有操作简单、易于并行处理、鲁棒性强等特点[6]，可以有效解决各种优化问题，并能够成功应用到聚类问题中提高算法的准确性,受到了很多学者的关注和研究。针对 K-means 算法存在过度依赖初始聚类中心而陷入局部最优的问题，潘晓英等人提出了基于自适应步长的萤火虫划分聚类算法，采用了自适应步长代替原有的固定步长[7]。王冲等人提出一种新的小生境萤火虫划分聚类算法以增加种群多样性[8]。陈小雪等人利用萤火虫优化算法全局搜索能力强、易于实现的优势，通过引用一种加权的欧氏距离对 K-means 算法的初始聚类中心进行优化[9]。上述基于萤火虫优化的 K-means 算法在给定聚类中心后的优化过程方面改善了聚类效果，但是都没有明确给出确定聚类中心数值 K 的方法，同时，基于自适应步长的改进算法在寻优后期设置的步长较小，易导致收敛速度较慢，且数据集局限于某个聚类中心无法跳出，从而陷入局部最优解，降低聚类精度。

收稿日期：2017-12-13；**修回日期：**2018-02-09　　**基金项目：**黑龙江省自然科学基金面上资助项目（F201422）

作者简介：杨明极（1971-），男，教授，博士研究生，主要研究方向为面向服务的软件工程、数据挖掘与离群检测（18846459670@163.com）；马池（1992-），女，硕士研究生，主要研究方向为聚类分析与数据挖掘；王娅（199-），女，硕士研究生，主要研究方向为数据挖掘；张竹（198-），女，博士研究生，主要研究方向为人工智能.

针对以上问题，本文提出一种基于混沌理论和最大最小距离算法的动态种群萤火虫算法(dynamic Firefly algorithm based on Chaos theory and Max-Min distance algorithm, FCMM)。该算法首先利用最大最小距离算法确定聚类中心数值 K 的大小和初始聚类中心位置,再用遍历性均匀、迭代速度快的 Tent 混沌映射构建以初始聚类中心为基准点的混沌搜索空间，并通过 Tent 混沌搜索的形式优化初始聚类中心，促使算法跳出局部最优，且可以获得较快收敛速度，最后利用智能萤火虫优化算法的位置更新公式划分非聚类中心样本点所属的聚类类别，完成聚类过程。

1　基本算法

1.1　K-means 聚类算法

K-means 的核心功能是将给定的数据集 $X=\{X_1,X_2,...,X_n\}$ 划分成 K 个子集 $\{C_1,C_2,...,C_k\}$，经典 K-means 聚类算法思想：从数据集 X 中随机选择 K 个对象，分别作为 K 个类别的初始聚类中心 $C_j(j=1,2,...,k)$；计算剩余每个对象 $X_i(i=1,2,...,n)$ 与各个聚类中心的欧氏距离，并将其划分到距离最近的子类 C_j 中；然后重新计算每个子类中所有对象的平均值，将其作为新的聚类中心[10]。重复上述过程，直到聚类中心不再改变。

定义 1 欧氏距离(Euclidean distance) 是欧氏空间中两点之间的直线距离[11]。样本 X_i 与 X_j 在 m 维空间中的欧式距离为

$$\mathrm{d}(X_\mathrm{i},X_\mathrm{j})=\sqrt{\sum_{\mathrm{k}=1}^{\mathrm{m}}(X_{\mathrm{ik}}-X_{\mathrm{jk}})^2} \tag{1}$$

K-means 算法根据欧式距离确定样本的相似程度。

1.2　萤火虫算法

该算法模拟萤火虫的移动过程，根据各萤火虫的位置和荧光亮度赋予其目标函数值并计算相对吸引度。荧光较强的个体吸引荧光较弱的个体按照位置更新公式向其移动，移动距离由吸引度的大小决定。优化过程基于以下三个原则：

a) 萤火虫的性别因素忽略，任何两只萤火虫个体均可以互相吸引。

b) 萤火虫个体的吸引度与距离成反比，与亮度成正比，亮度强的萤火虫吸引亮度弱的萤火虫向其移动，亮度最强的个体随机移动；。

c) 萤火虫个体的亮度由其所在位置的目标函数值决定。

定义 2　荧光亮度为

$$I\propto -f(x_i),1\le i\le n \tag{2}$$

$$I=I_0\exp(-\gamma * r_{ij}) \tag{3}$$

其中 I 表示荧光亮度，$f(x_i)$ 表示目标函数，x_i 为萤火虫 i 的空间位置；I_0 表示最大荧光亮度；γ 为常量，表示光强吸收因子；r_{ij} 表示 x_i 与 x_j 之间的欧式距离。

定义 3　吸引度为

$$\beta=\beta_0\exp(-\gamma * r_{ij}^2) \tag{4}$$

其中，β_0 为最大吸引度。

定义 4　位置更新公式为

$$x_i(t+1)=x_i(t)+\beta(x_j(t)-x_i(t))+\alpha\varepsilon_i \tag{5}$$

其中:α 表示初始化的步长因子；ε_i 表示服从高斯分布的随机因子。

2　K-means 聚类算法的优化

对于传统 K-means 算法，给定初始聚类中心后，可以利用萤火虫算法的随机性和全局搜索能力对除去聚类中心以外的数据集进行较为精确的划分，进而提高算法的收敛速度，实现对 K-means 算法的优化。但是基于萤火虫优化的 K-means 算法存在以下缺陷：

a)聚类中心数值 K 的选取尚没有指定算法，若选取的 K 值不合理，会严重影响聚类精度和计算复杂度。

b)最佳聚类结果对应于目标函数的极值点，聚类中心落在某个局部极小点附近，易导致算法陷入局部最优[10]。

在利用萤火虫个体的相互吸引模拟 K-means 聚类过程，提高全局搜索收敛速度的基础上，针对现有算法在全局寻优搜索中易陷入局部极值区域的缺陷，本文根据混沌映射对初值具有敏感性和遍历性的特点[12]，提出一种基于混沌理论的动态种群萤火虫算法(简称 FCMM 算法)，对 K-means 聚类进行改进。

2.1　最大最小距离算法确定聚类中心数值 K

最大最小距离(Max-Min distance)聚类算法，本文简称 MM 算法。MM 算法与传统 K-means 相似，都是通过计算欧氏距离，根据最近邻原则划分归属于各聚类中心的样本点。不同之处在于，MM 算法并不是直接给定聚类类别数值 K，而是从样本点中任选一个对象 X_i，作为第一个聚类中心，通过式(1)计算各点到 X_i 的欧式距离，将距离 X_i 最远的点作为新的聚类中心。重复以上划分步骤，直到不再产生新的聚类中心，最后确定聚类中心总数 K。

算法步骤如下：

a)给定 $\theta,0<\theta<1$,选取初始聚类中心 $Z_1=x_1$。

b)产生新的聚类中心

通过计算各点到 Z_1 的欧氏距离 D_{i1}，选取 $D_{k1}=\max\{D_{i1}\}$ 对应的 x_k 为下一个聚类中心 Z_2；

计算各点到聚类中心 Z_1 和 Z_2 的距离 D_{i1},D_{i2}，若 $D_l=\max\{\min(D_{i1},D_{i2})\},i=1,2,...n$, 并且 $D_l>\theta * D_{12}$,，则取 x_l 为第三个聚类中心 Z_3。

$$D_{i1}=\|x_i-Z_1\|=\sqrt{\sum_{i=1}^{d}|x_i-Z_1|^2},D_{i2}=\|x_i-Z_2\|$$

注：D_{12} 表示 Z_1 和 Z_2 之间的距离

若 Z_3 存在，判断是否有 $D_j=\max\{\min(D_{i1},D_{i2},D_{i3})\}$, $i=1,2,...n$, 若满足以上条件并且 $D_l>\theta * D_{12}$, 则确定第四个聚类中心。依次类推，如果出现 $D_l\le\theta * D_{12}$,，停止寻找新的聚类中心。

c)统计聚类中心总数 K。

该算法的聚类结果与参数和起始点的选取关系重大。为得

到良好的聚类效果，在无先验样本分布知识的情况下需要进行反复实验，故此算法在本文仅用于确定聚类中心数值 K 的大小。

2.2 混沌理论优化聚类中心

为避免基于萤火虫优化的 K-means 算法的聚类中心落在某个局部极小点附近，改善算法的全局搜索能力，尽快跳出局部最优解，本文引入具有随机性、遍历性、规律性特征的混沌变量，对基于萤火虫算法的 K-means 聚类进行优化处理。

由于 Logistic 混沌映射在 0,1 边界范围分布集中，存在明显的遍历不均匀的问题，导致算法效率较低。单梁[10]等通过严谨的推理证明，验证了 Tent 映射产生的混沌序列更有助于算法优化[13]，并指出 Tent 映射的收敛速度和遍历均匀性等性能相较于 logistic 映射更优。

2.2.1 Tent 混沌序列

Tent 映射表达式为

$$x_{t+1}=\begin{cases}2x_t,0\le x_t\le\frac{1}{2}\\ 2(1-x_t),\frac{1}{2}\le x_t\le 1\end{cases} \tag{6}$$

Tent 映射经伯努利移位变换后表示为

$$x_{t+1}=(2x_t)\mathrm{mod}1 \tag{7}$$

Tent 混沌序列的产生步骤如下：

a)随机产生一个不在(0.20,0.40,0.60,0.80) 范围内的初值 x_0，记作 $z,z(1)=x_0,i=j=1$；

b)按式(7)进行迭代，产生序列 x 。

c) 如果 $x(i)=[0,0.25,0.5,0.75]$ 或 $x(i)=x(i-k)$, $k=[0,1,2,3,4]$,执行步骤 b)。

d)按式 $x(i)=z(j+1)$ 改变迭代初值，j=j+1，执行步骤 2。

e)如果达到最大迭代次数，终止运行，保存产生的 x 序列。

2.2.2 混沌搜索

本文提出的 FCMM 算法以当前搜索到的局部最优解为基准点产生 Tent 混沌序列，通过 Tent 搜索跳出局部最优，从而获得全局最优解。

具体方案如下：将各个聚类中心 $C_i(i=1,2,...k)$ 与当前聚类中心 C_x 的距离 D_{ix} 从大到小依次排列，取距离较小的前面 n 个(占聚类中心总数的 30%)类 $C_{i1},C_{i2},...,C_{in}$ 与 C_x，并分别求出当前 n+1 个类中第 j 维的最大值 $X^j_{\max}$，最小值 $X^j_{\min}$，共同构成新的混沌搜索空间。以 C_x 的聚类中心 X_x 为基准点产生混沌序列，进行混沌搜索，搜索结束得到的最优解作为新的聚类中心。

假设聚类中心是 C_x， $X_k=\{x_{k1},x_{k2},...x_{kd}\}$,

$$x_{kj}\in[X_{\min},X_{\max}]。$$

Tent 混沌搜索主要步骤：

a)利用 $z^0_{kj}=(x_{kj}-X^j_{\min})/(X^j_{\max}-X^j_{\min})$,将 X_x 映射到(0,1)，其中 $k=1,2,...,n,j=1,2,...,D$。

b)将上式代入式(7)进行 Tent 映射，迭代产生混沌变量序列 $z^m_{kj}(m=1,2,...,C_{\max})$。 $C_{\max}$ 是混沌搜索的最大迭代次数[11]。

c)利用式(8)将 z^m_{kj} 还原到原解空间的邻域内，产生新解 V_k。

$$v_{kj}=x_{kj}+\left(X^j_{\max}-X^j_{\min}\right)*\left(2z^m_{kj}-1\right)/2 \tag{8}$$

d)计算 v_k 的荧光亮度值 $F(v_k)$，并与局部最优解的荧光亮度值 $F(x_k)$ 比较，保留最好的解。

e)若搜索次数达到 $C_{\max}$，停止搜索；否则，转向步骤 2。

2.3 FCMM 算法基本步骤

a)初始化参数：聚类对象总数 N,吸收系数 γ，步长因子 α，混沌搜索的最大迭代次数 $C_{\max}$，最大荧光亮度 I，最大吸引度 β_0。

b)通过最大最小距离算法确定聚类中心数 K，记录最大最小距离算法获得的初始聚类中心位置。

c)依次通过 Tent 映射构建以各个聚类中心为基准点的混沌搜索空间。

d)利用 Tent 混沌搜索更新初始聚类中心的位置，直到聚类中心不再变化。

e)将聚类中心对应于目标萤火虫，赋予最高荧光亮度。计算剩余的样本点相对于各聚类中心的欧氏距离，并按照式(3)赋予不同的荧光亮度。

f)如果 $I_i>I_j$,表示萤火虫 j 比 i 的目标函数值小，即 j 比 i 的位置好，萤火虫 j 将吸引 i 向它移动，移动方式由式(4)决定，并通过式(5)更新萤火虫位置。

g)重复步骤 f)，直到所有萤火虫都被划分到所属的聚类中心。

h)输出结果。

3 实验结果与分析

3.1 实验环境

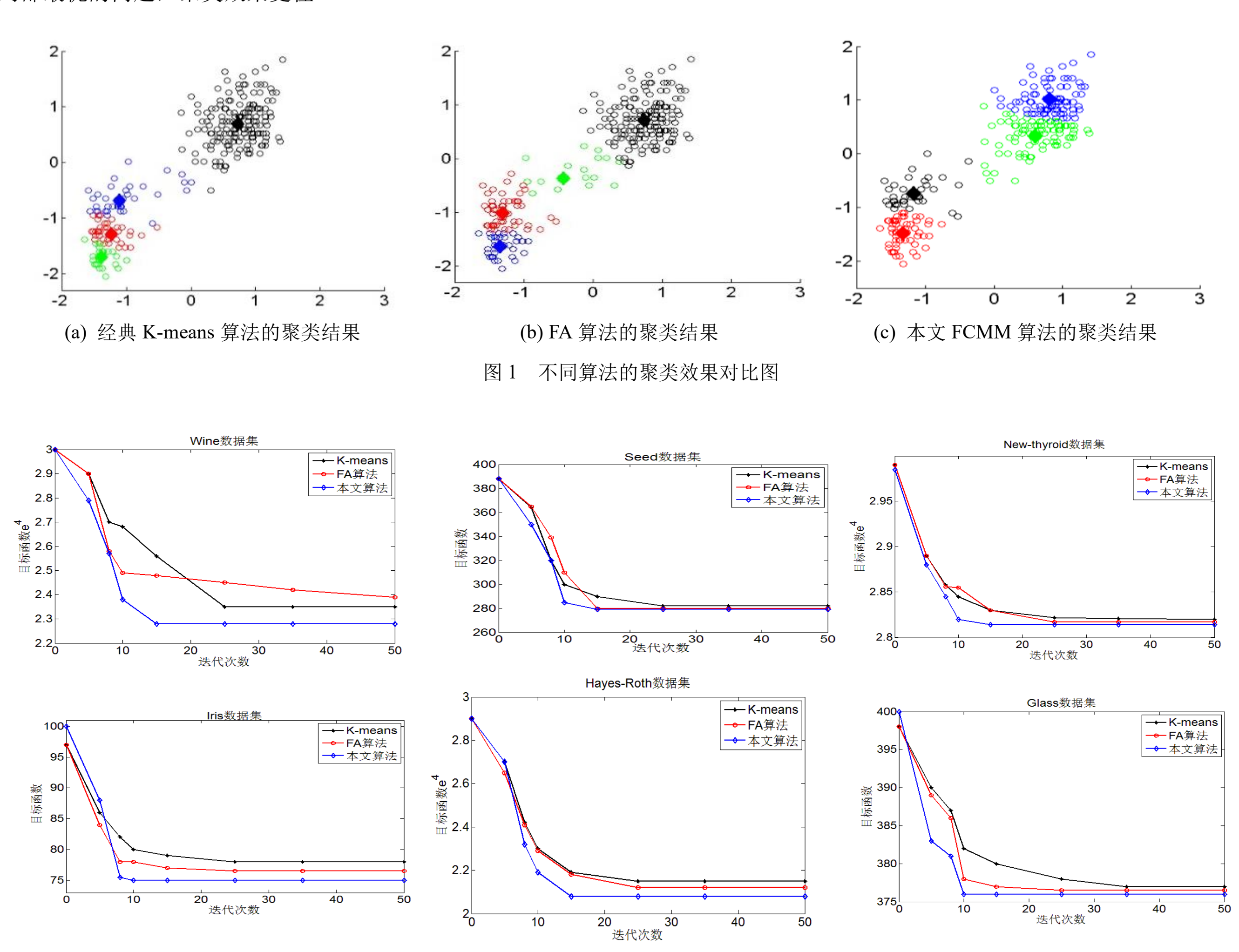
为了验证算法的有效性，本文进行了三组实验。实验一通过对比不同算法的聚类效果图，验证本文算法聚类中心选取的有效性；实验二通过 UCI 数据集测试不同聚类算法的聚类精度和收敛速度，验证本文算法收敛速度较快、聚类精度有一定提高。实验三将本文算法与引言中介绍的基于自适应步长的萤火虫划分聚类算法（以下简称文献[7]算法）和加权的欧氏距离对 K-means 改进算法（以下简称文献[9]算法）的聚类效果进行对比分析。实验的运行环境为 Windows7 操作系统，4 GB 物理内存，CPU 速度 3.10 GHz，matlab2014b。

3.2 实验结果与分析

实验 1　聚类效果对比。随机选取 200 个样本数据散布在解空间，经过最大最小距离算法得到聚类类别数 K=4。鉴于萤火虫算法的参数设定对实验结果有很大影响，本文采用水平实验得到最优解出现次数最多的组合，对步长因子 α 和光强吸收系数 γ 采用枚举法得到一系列最优值组合，最后对结果进行分析得到了 FA 算法的参数取值情况。经 30 组实验结果分析对比，参数设置如下都会取得比较好的效果：最大吸引度 β_0 =100，吸收系数 γ =1，步长因子 α =0.06，最大迭代次数 $C_{\max}$ =50，最大荧光亮度 I =100， $\theta=0.4$。测试结果如图 1 所示。

由图 1 可以看出，传统 K-means 算法的部分聚类中心分布较为集中，明显存在局部最优问题；FA 算法获得的聚类中心分布均匀性稍有改善，但聚类效果仍有待提高；本文算法相较于 K-means 和 FA 算法，聚类中心分布更均匀，明显改善了易陷入局部最优的问题，聚类效果更佳。

实验 2　选用标准数据集进行仿真，参数设置同实验一。为了验证本文算法的有效性，K-means、FA 算法与 FCMM 算法分别在 6 种不同的数据集上进行独立实验，不同算法的平均聚类精度如表 1 所收敛曲线如图 2 所示。



(a) 经典 K-means 算法的聚类结果　(b) FA 算法的聚类结果　(c) 本文 FCMM 算法的聚类结果

图 1　不同算法的聚类效果对比图

图 2　三种算法在 6 种数据集上的收敛曲线

表 1　算法的平均聚类精度(%)

数据集	K-means	FA 算法	本文算法
Iris	87.93	91.13	92.16
Wine	56.85	70.23	72.15
Seed	86.97	88.07	90.46
Glass	54.05	57.18	63.12
Hayes-Roth	77.32	81.06	82.35
New-thyroid	72.34	79.63	80.28

由表 1 可以看出，本文算法首先通过最大最小距离算法确定聚类中心数值 K 并引入混沌理论对聚类中心进行优化后，平均聚类精度相较于经典 K-means 算法和 FA 算法分别提高了 7.51%和 2.2%。由图 2 看出，本文算法在保证较高聚类精度的前提下，相比于经典 K-means 和 FA 算法有较快的收敛速度。

实验 3　不同算法聚类效果对比分析。该实验将本文算法与引言中的基于萤火虫优化的 K-means 算法在聚类准确率和处理时间等方面进行对比分析。

由表 2 可知，相对于文献[7]与文献[9]的算法，本文算法由于采用了最大最小距离算法进行初始聚类中心数值的分析，时间复杂度相对较高，处理时间略长；但是聚类精度明显较高。

4　结束语

本文针对 K-means 算法易受初始聚类中心影响而陷入局部最优的问题，提出了新的 FCMM 聚类算法，分别在初始聚类中心的个数和聚类中心的位置分布两个方面对 K-means 算法进行了改进。本文提出的确定聚类中心数值 K 的方式，可以有效降低算法受初始聚类中心个数的影响。同时，在基于萤火虫优化 K-means 聚类算法的基础上，提出了通过混沌搜索的方式更新聚类中心位置的方法，由于混沌映射可以降低初始聚类位置对结果的影响，同时充分利用了萤火虫算法的寻优能力和收敛速度，成功解决了聚类过程易陷入局部最优解和收敛速度过慢的问题。

表 2　不同算法的仿真结果比较

算法	本文算法			文献[7]算法			文献[9]算法		
数据集	Glass	Hayes-Roth	Wine	Glass	Hayes-Roth	Wine	Glass	Hayes-Roth	Wine
聚类精度/%	63.12	82.35	72.15	62.45	81.28	70.71	60.70	81.46	70.43
处理时间/s	19.425	11.432	15.725	17.532	10.734	13.832	18.345	11.243	14.432

算法	本文算法			文献[7]算法			文献[9]算法		
数据集	Iris	New-thyroid	Seed	Iris	New-thyroid	Seed	Iris	New-thyroid	Seed
聚类精度/%	92.16	80.28	90.46	91.22	79.01	89.52	91.49	79.98	89.70
处理时间/s	10.003	12.709	10.782	9.234	10.232	8.342	8.342	11.344	9.232

经过验证不同算法聚类效果的对比和UCI数据集聚类结果的分析，验证了本文算法在对少量数据进行聚类分析的过程中具有收敛速度快，聚类精度高，不易陷入局部最优问题。但是使用本文提出的算法处理较大数据集时，会产生时间复杂度较高的问题，所以需要研究如何更好的使用相似性准则去掉聚类中心候选集，这也是本文算法今后的改进和研究方向。

参考文献：

[1] Borgwardt S, Brieden A, Gritzmann P. An LP-based K-means algorithm for balancing weighted point sets [J]. European Journal of Operational Research, 2017, 263 (2) .

[2] Wangchamhan T, Chiewchanwattana S, Sunat K, *et al.* Efficient algorithms based on the K-means and Chaotic League Championship Algorithm for numeric, categorical, and mixed-type data clustering [J]. Expert Systems with Applications, 2017, 90.

[3] Patel G K, Dabhi V K, Prajapati H B. Clustering using a combination of particle swarm optimization and K-means [J]. Journal of Intelligent Systems, 2016, 26 (3) .

[4] 陈兴蜀，吴小松，王文贤，等. 基于特征关联度的 K-means 初始聚类中心优化算法 [J]：四川大学学报：工程科学版, 2015, 47 (1): 13-19.

[5] 张强，王红卫，陈游，等. 基于自适应权重的 RFCM 聚类算法 [J]. 微电子学与计算机, 2016, 33 (12): 80-84.

[6] 赵杰. 群智能优化算法在聚类分析中的应用研究 [D]. 西安：陕西师范大学, 2015.

[7] 潘晓英，陈雪静，李昂儒，等. 基于自适应步长的萤火虫划分聚类算法 [J]. 计算机应用研究, 2017, 34 (12): 3576-3579.

[8] 王冲，雷秀娟. 新的小生境萤火虫划分聚类算法 [J]. 计算机工程, 2014, 40 (5): 173-177.

[9] 陈小雪，尉永清，任敏，等. 基于萤火虫优化的加权 K-means 算法 [J]. 计算机应用研究, 2018, 35 (2) .

[10] 陈望，贾振红，覃锡忠，等. 基于改进 K-means 聚类算法的室内 WLAN 定位研究 [J]. 激光杂志, 2014, 35 (7): 11-14.

[11] 徐鹏程，王诚. K-means 算法改进及基于 Spark 计算模型的实现 [J]. 南京邮电大学学报：自然科学版, 2017, 37 (04): 113-118.

[12] 匡芳君，徐蔚鸿，金忠. 自适应 Tent 混沌搜索的人工蜂群算法 [J]. 控制理论与应用, 2014, 31 (11): 1502-1509.

[13] 单梁，强浩，李军，等. 基于 Tent 映射的混沌优化算法 [J]. 控制与决策, 2005, 20 (2): 179-182.

[14] 赵杰，雷秀娟，吴振强，等. 基于最优类中心扰动的萤火虫聚类算法 [J]. 计算机工程与科学, 2015, 37 (2): 342-347.